CLAIMS

What is claimed is:

1. A method of searching a database containing hypertext documents, said method comprising:

searching said database using a query to produce a set of hypertext documents; and

clustering said set of hypertext documents into various clusters such that documents within each cluster are similar to each other.

2. The method in claim 1, wherein said set of hypertext documents comprises a collection of unstructured, unlabeled documents and said clustering organizes said set of hypertext documents into labeled categories that are discriminated and disambiguated from each other.

3. The method in claim 1, wherein said clustering is based upon words contained in each hypertext document, out-links from each hypertext document, and in-links to each hypertext document.

4. The method in claim 3, wherein said hypertext documents are considered similar if said hypertext documents share one or more of said words, said out-links, and said in-links.

5. The method in claim 3, wherein said clustering includes determining a relative importance of said words, said out-links, and said in-links in an adaptive, data-driven process.

6. The method in claim 1, further comprising annotating each cluster using information nuggets.

7. The method in claim 6, wherein said information nuggets include nuggets relating to summary, breakthrough, review, keywords, citation, and reference.

8. The method in claim 7, wherein said summary and said keywords are derived from said words, said review and said references are derived from said out-links, and said breakthrough and said citations are derived from said in-links.

9. The method in claim 7, wherein said summary comprises a document in a cluster having a most typical in-link feature vector amongst all documents in said cluster.

10. The method in claim 7, wherein said breakthrough comprises a document in a cluster having a most typical in-link feature vector amongst all documents in said cluster.

11. The method in claim 7, wherein said review comprises a document in a cluster having a most typical out-link feature vector amongst all documents in said cluster.

12. The method in claim 7, wherein said keyword comprises a word in said word dictionary for said cluster that has a largest weight.

13. The method in claim 7, wherein said citation comprises a document in a cluster representing a most typical in-link into said cluster.

14. The method in claim 7, wherein said reference comprises a document in a cluster representing a most typical out-link out of said cluster.

15. A method of searching a database containing hypertext documents, said method comprising:

searching said database using a query to produce a set of hypertext documents; and

clustering said set of hypertext documents into various clusters such that documents within each cluster are similar to each other,

wherein said clustering is based upon words contained in each hypertext document, out-links from each hypertext document, and in-links to each hypertext document.

16. The method in claim 15, wherein said set of hypertext documents comprises a collection of unstructured, unlabeled documents and said clustering organizes said set of hypertext documents into labeled categories that are discriminated and disambiguated from each other

17. The method in claim 15, wherein said hypertext documents are considered similar if said hypertext documents share one or more of said words, said out-links, and said in-links.

18. The method in claim 15, wherein said clustering includes determining a relative importance of said words, said out-links, and said in-links in an adaptive, data-driven process.

19. The method in claim 15, further comprising annotating each cluster using information nuggets.

20. The method in claim 19, wherein said information nuggets include nuggets relating to summary, breakthrough, review, keywords, citation, and reference.

21. The method in claim 20, wherein said summary and said keywords are derived from said words, said review and said references are derived from said out-links, and said breakthrough and said citations are derived from said in-links.

22. The method in claim 20, wherein said summary comprises a document in a cluster having a most typical in-link feature vector amongst documents in said cluster.

23. The method in claim 20, wherein said breakthrough comprises a document in a cluster having a most typical in-link feature vector amongst documents in said cluster.

24. The method in claim 20, wherein said review comprises a document in a cluster having a most typical out-link feature vector amongst documents in said cluster.

25. The method in claim 20, wherein said keyword comprises a word in said word dictionary for said cluster that has a largest weight.

26. The method in claim 20, wherein said citation comprises a document in a cluster representing a most typical in-link into said cluster.

27. The method in claim 20, wherein said reference comprises a document in a cluster representing a most typical out-link out of said cluster.

28. A method of searching a database of documents comprising:

performing a search of said database using a query to produce query result documents;

constructing a word dictionary of words within said query result documents;

constructing an out-link dictionary of documents within said database that are pointed to by said query result documents;

adding said query result documents to said out-link dictionary;

constructing an in-link dictionary of documents within said database that point to said query result documents; and

adding said query result documents to said in-link dictionary.

29. The method in claim 28, further comprising:

forming first vectors for words remaining in said word dictionary;

forming second vectors for documents remaining in said out-link dictionary;

forming third vectors for documents remaining in said in-link dictionary;

normalizing said first vectors, said second vectors, and said third vectors to create vector triplets for document remaining in said in-link dictionary and said out-link dictionary; and

clustering the said vector triplets into one of clusters, classes and partitions.

30. The method in claim 29, where said clustering comprises a four step *toric k-means* process comprising:

(a) arbitrarily segregating the vector triplets into clusters;

(b) for each cluster, computing a set of concept triplets describing said cluster;

(c) re-segregating said vector triplets into a more coherent set of clusters by putting each vector triplet into a cluster corresponding to a concept triplet that is most similar to, a given vector triplet; and

(d) determining a coherence for each of said clusters based on a similarity of vector triplets within each of said clusters, and repeating steps (b)-(c) until coherence of the obtained clusters no longer significantly increases.

31. The method in claim 29, further comprising annotating and summarizing said clusters using nuggets of information, said nuggets including summary, breakthrough, review, keyword, citation, and reference.

32. The method in claim 31, wherein said summary comprises a document in a cluster having a most typical in-link feature vector amongst all documents in said cluster.

33. The method in claim 31, wherein said breakthrough comprises a document in a cluster having a most typical in-link feature vector amongst all documents in said cluster.

34. The method in claim 31, wherein said review comprises a document in a cluster having a most typical out-link feature vector amongst all documents in said cluster.

35. The method in claim 31, wherein said keyword comprises a word in said word dictionary for said cluster that has a largest weight.

36. The method in claim 31, wherein said citation comprises a document in a cluster representing a most typical in-link into said cluster.

37. The method in claim 31, wherein said reference comprises a document in a cluster representing a most typical out-link out of said cluster

38. The method in claim 28, further comprising pruning function words from said word dictionary.

39. The method in claim 28, further comprising pruning documents from said out-link dictionary that are pointed to by fewer than a first predetermined number of said query result documents.

40. The method in claim 28, further comprising pruning documents from said in-link dictionary that point to fewer than a second predetermined number of said query result documents.

41. A method of searching a database of documents comprising:

performing a search of said database using a query to produce query result documents;

constructing a word dictionary of words within said query result documents;

pruning function words from said word dictionary;

forming first vectors for words remaining in said word dictionary;

constructing an out-link dictionary of documents within said database that are pointed to by said query result documents;

adding said query result documents to said out-link dictionary;

pruning documents from said out-link dictionary that are pointed to by fewer than a first predetermined number of said query result documents;

forming second vectors for documents remaining in said out-link dictionary;

constructing an in-link dictionary of documents within said database that point to said query result documents;

adding said query result documents to said in-link dictionary;

pruning documents from said in-link dictionary that point to fewer than a second predetermined number of said query result documents;

forming third vectors for documents remaining in said in-link dictionary;

normalizing said first vectors, said second vectors, and said third vectors to create vector triplets for document remaining in said in-link dictionary and said out-link dictionary;

clustering the said vector triplets using a four step process of *toric k-means* comprising:

(a) arbitrarily segregating said vector triplets into clusters;

(b) for each cluster, computing a set of concept triplets describing said cluster;

(c) re-segregating said vector triplets into more coherent set of clusters by putting each vector triplet into a cluster corresponding to a concept triplet that is most similar to, a given vector triplet; and

(d) determining a coherence for each of said clusters based on a similarity of vector triplets within each of said clusters, and repeating steps (b)-(c) until coherence of the obtained clusters no longer significantly increases; and

annotating and summarizing said vector triplets using nuggets of information, said nuggets including summary, breakthrough, review, keyword, citation, and reference.

42. A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform a method of searching a database containing hypertext documents, said method comprising:

searching said database using a query to produce a set of hypertext documents; and

clustering said set of hypertext documents into various clusters such that documents within each cluster are similar to each other,

wherein said clustering is based upon words contained in each hypertext document, out-links from each hypertext document, and in-links to each hypertext document.

43. The program storage device in claim 42, wherein said set of hypertext documents comprises a collection of unstructured, unlabeled documents and said clustering organizes said set of hypertext documents into labeled categories that are discriminated and disambiguated from each other

44. The program storage device in claim 42, wherein said hypertext documents are considered similar if said hypertext documents share one or more of said words, said out-links, and said in-links.

45. The program storage device in claim 42, wherein said clustering includes determining a relative importance of said words, said out-links, and said in-links in an adaptive, data-driven process.

46. The program storage device in claim 42, further comprising annotating each cluster using information nuggets.

47. The program storage device in claim 46, wherein said information nuggets include nuggets relating to summary, breakthrough, review, keywords, citation, and reference.

48. The program storage device in claim 47, wherein said summary and said keywords are derived from said words, said review and said references are derived from said out-links, and said breakthrough and said citations are derived from said in-links.

49. The program storage device in claim 47, wherein said summary comprises a document in a cluster having a most typical in-link feature vector amongst documents in said cluster.

50. The program storage device in claim 47, wherein said breakthrough comprises a document in a cluster having a most typical in-link feature vector amongst documents in said cluster.

51. The program storage device in claim 47, wherein said review comprises a document in a cluster having a most typical out-link feature vector amongst the documents in said cluster.

52. The program storage device in claim 47, wherein said keyword comprises a word in said word dictionary for said cluster that has a largest weight.

53. The program storage device in claim 47, wherein said citation comprises a document in a cluster representing a most typical in-link into said cluster.

54. The program storage device in claim 47, wherein said reference comprises a document in a cluster representing a most typical out-link out of said cluster